Nor Jnl Ling 39.1, 5–37 © Nordic Association of Linguists 2016 doi:10.1017/S0332586515000293

Grigonytė, Gintarė, Maria Kvist, Mats Wirén, Sumithra Velupillai & Aron Henriksson. 2016. Swedification patterns of Latin and Greek affixes in clinical text. *Nordic Journal of Linguistics* 39(1), 5–37.

Swedification patterns of Latin and Greek affixes in clinical text

Gintarė Grigonytė, Maria Kvist, Mats Wirén, Sumithra Velupillai & Aron Henriksson

Swedish medical language is rich with Latin and Greek terminology which has undergone a Swedification since the 1980s. However, many original expressions are still used by clinical professionals. The goal of this study is to obtain precise quantitative measures of how the foreign terminology is manifested in Swedish clinical text. To this end, we explore the use of Latin and Greek affixes in Swedish medical texts in three genres: clinical text, scientific medical text and online medical information for laypersons. More specifically, we use frequency lists derived from tokenised Swedish medical corpora in the three domains, and extract word pairs belonging to types that display both the original and Swedified spellings. We describe six distinct patterns explaining the variation in the usage of Latin and Greek affixes in clinical text. The results show that to a large extent affixes in clinical text are Swedified and that prefixes are used more conservatively than suffixes.

Keywords affixes, clinical text, corpus linguistics, health records, Latin and Greek terminology

Gintarė Grigonytė, Department of Linguistics, Stockholm University, 106 91 Stockholm, Sweden. gintare@ling.su.se
Maria Kvist, Department of Computer and Systems Sciences, Stockholm University, Postbox 7003, 164 07 Kista, Sweden. maria.kvist@karolinska.se
Mats Wirén, Department of Linguistics, Stockholm University, 106 91 Stockholm, Sweden. mats.wiren@ling.su.se
Sumithra Velupillai, Department of Computer and Systems Sciences, Stockholm University, Postbox 7003, 164 07 Kista, Sweden. sumithra@dsv.su.se
Aron Henriksson, Department of Computer and Systems Sciences, Stockholm University, Postbox 7003, 164 07 Kista, Sweden. aronhen@dsv.su.se

1. INTRODUCTION

Medical terminology in Germanic and other languages has a large stock of Latin and Greek prefixes, roots and suffixes. By and large, Greek is the language of pathology (the study of diseases) and Latin is the language of anatomy (the structure of the body). In Swedish medical language, two parallel developments can be seen with respect to this terminology. On the one hand, according to Nyman (2013a:43), the

overall use of Latin and Greek terms in medical language appears to have increased since the 1950s. For example, Swedish terms that were common 50–60 years ago, such as *sockersjuka*, literally: 'sugar disease', *barnförlamning* 'children's palsy' and *kräfta* 'crayfish', are nowadays replaced by *diabetes*, *polio* and *cancer*. It seems that there are several reasons for this: Latin and Greek terms are precise, largely void of expressive meaning, easily adaptable into Swedish linguistic patterns, and often have direct correspondences in English and other languages. Even spin-offs of these kinds of terms into the general language are gaining ground, for example, 'traffic infarct' and 'corporate anorexia'. On the other hand, the Health Record Act (*patientjournallagen*), adopted in 1985, brought about the first regulation on Swedification and standardization of foreign medical vocabulary, motivated by a demand for transparency and patient empowerment. SWEDIFICATION (*försvenskning*) here means adaptation to Swedish spelling and inflection. This can be contrasted with translation, which means forming an equivalent using Swedish vocabulary. For example, Swedification of *bronchitis* gives *bronkit*, whereas translation gives *luftrörskatarr* (Fogelberg & Petersson 2013:12).

Although spelling of Latin and Greek vocabulary according to Swedish conventions was regulated in 1987 (Smedby 1991, 2013:185), adherence to this in the medical community has not been univocal. As a result, the overall spelling variation has rather increased, with differences depending on medical profession, medical domain, and also the kind of Latin and Greek morphemes involved. In addition to this, there is a strong influence from the English spelling of Latin and Greek, resulting in a mixture of Latin, Greek, English and Swedish spellings, sometimes in the same word. The combinatorics of these influences is enormous, giving rise to huge numbers of spelling variants of the same terms (Grigonytė et al. 2014). This in turn constitutes a problem for laypersons trying to look up terms from clinical text and a serious obstacle for automatic language processing for the purpose of simplification, normalization and text mining.

This paper is a case study in the terminological variation in the domain of health records. A health record contains systematic documentation of a single patient's medical history across time, entered by healthcare professionals with the purpose of enabling informed care. The language in this domain, which we refer to as clinical text, is produced by people who are, on the one hand, highly specialised professionals, but on the other hand are non-professional writers, giving rise to a genre which is highly interesting from a linguistic viewpoint. The goal of the study is to obtain precise quantitative measures of how the foreign terminology is manifested in Swedish clinical text. To this end, we shall study the effects of spelling influences along three dimensions: different Latin and Greek prefixes and suffixes, different medical professions, and different medical subspecialties. The rationale for confining the study to prefixes and suffixes is that the behaviour of these can be exhaustively analysed since they constitute a closed class of morphemes; at the same time, they combine

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

with different stems in a highly productive way. As baselines for comparison, we use general medical language from a Swedish medical journal (*Läkartidningen*) and from a public website dedicated to medical counselling (Vårdguiden).

The purpose of the study is to answer the following research questions: How far has the process of Swedification come in the 20 years since the Health Record Act? Has the effect been to actually increase the number of spelling variants instead of standardising them? Are prefixes or suffixes more resistant to Swedification? Do linguistic factors such as position of affixes inside a word, or external factors such as domain or profession of the language user, play a role in the extent to which Swedification progresses? From a theoretical point of view, these questions are related to morphological connectivity and the combinatorial properties of affixes (Hay & Plag 2004, Baayen 2010). From a descriptive point of view, this study can be seen as an elaboration of Fogelberg & Petersson (2013), with qualitative and quantitative detail for several medical genres and types of language users. In addition, the results bear on the development of computational methods for processing of medical text for purposes such as normalization of vocabulary or information retrieval.

2. BACKGROUND

2.1 Swedish clinical text

Medical text as it is found in textbooks and journals, on the one hand, and the language of health records, on the other, are written under different conditions and for different purposes, and therefore differ substantially (Friedman, Kra & Rzhetsky 2002, Smith et al. 2014). The purpose of medical text is to transfer knowledge, which requires formal, well-structured and correct text, whereas health records are written under time pressure, being used as memory notes or information for the professional team, and seldom corrected by the author. In both of these domains, medical terminology is used to convey information as precisely and concisely as possible; in the case of health records, a key purpose of this is to ensure patient safety.

2.1.1 History and legislation of patient records

To give some context to medical documentation and its history from the perspective of how clinical notes are written, we provide a brief historical and legislative outline. Medical records have been kept in Sweden at least since the 18th century. A 1730 thesis, written in Latin, *Historiis moriborum rite consignandis* by the Swedish physician Nils Rosén von Rosenstein, states that the purpose of keeping records by doctors is not only to be of use in the care of a patient but also to accumulate knowledge, in line with Hippocrates' thoughts (Nilsson 2007). The author also states criteria for the content and structure of the patient record. These early records were

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

mostly written in Latin with the Greek words of pathology. In 1863, the economic logic started to influence the medical content as rules and regulations stated that the recording of the number of operations, hospitalizations and clinical visits was the base for reimbursement. This influence of economic reimbursement is still strong in the construction of electronic health records systems. During the 20th century, the medical record also became a legal document, and as the legal rights of the patients was regulated this would also influence the clinical texts. In Sweden, the habit of suing the doctor for malpractice is not at all as spread as in the USA, for example, but the eminent threat does influence medical professionals further to be thorough and precise in their documentation of given care. For this purpose also, the need of a precise medical terminology is evident.

Today, the documentation is regulated by the Patient Data Act (Socialdepartementet 2008). It is stated that the foremost purpose of patient record documentation is to contribute to good and safe healthcare (Patient Data Act, Chapter 3, §2). However, the legislation also regulates the language to be used in patient records, and has since 1985 included a directive on Swedish as the preferred language. The decree that the records should be written in a language that is comprehensible for the patient has never really been given preference among physicians, as they foremost see the records as a working tool for the professionals, and prioritize the main purpose of safe healthcare (Allvin 2010), hence the use of technical terminology is heavy.

2.1.2 Characterization of clinical language in electronic patient records

The process of transferring patient records from paper documents into electronic records has made it possible to study and develop natural language processing (NLP) tools for information extraction and other useful methods and tools. However, since health records are sensitive texts and protected by confidentiality, the availability of large corpora for scientific studies, for example linguistic studies, is still limited.

The transfer to the electronic media has not led to improvements of the clinical texts in records as much as could have been expected. The possibilities of using the textual documentation for e.g. visualization of clinical events in timelines, tables or other graphs have been surprisingly unexplored. Also, automated documentation support such as free text search, spelling and language checking, and summarization – functions that are common in other documentation systems – has not been applied for health record systems to any greater extent as of yet. The opportunity to transfer data into more structured records, thus enabling automatic functions and statistical evaluation of health care has been explored in some health record systems, for some types of information, but much of the documentation is written as it always has been; in unstructured free-text paragraphs rich in medical terminology.

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

Characteristics of clinical text are surprisingly similar in different even unrelated languages (Friedman et al. 2002, Surján & Héja 2003, Laippala et al. 2009, Hagège et al. 2011, Bretschneider, Zillner & Hammon 2013, Temnikova et al. 2013, Smith et al. 2014). Several of these characteristics reflect the constant time pressure in healthcare, such as telegraphic text omitting words and frequent use of ad hoc abbreviations. Also, the fact that many physicians use dictaphones for documentation may sometimes contribute to an unusual sentence structure, containing many subordinate clauses – but this is less frequent. Most sentences in health records are very short (less than 11 words on average) and are not transcribed (Smith et al. 2014). Clinical text is heavy with technical terms, many originating from Latin, Greek or English. The nature of the diagnosis process results in many negated or speculative statements. The omission of subjects leads to information-dense sentences. Moreover, earlier studies of Swedish clinical text report frequent use of verb less sentences, i.e. 63% of sentences in a corpus of radiology reports lacked a main verb (Smith et al. 2014). This is in line with findings in German and Bulgarian clinical text (Bretschneider et al. 2013, Temnikova et al. 2013).

2.1.3 Clinical subdomains and domain language

There are differences between subdomains of clinical text, e.g. different language use by different healthcare professions, in part owing to different vocabulary due to their diverse chores but also due to varying academic training. Other variations in language use can be seen between subspecialties within the clinical professions (Patterson & Hurdle 2011, Zeng et al. 2011), not only because of different working conditions and tasks, but also on the account of the varying cultures. During medical profession education and training, emphasis on teaching and learning about healthcare documentation lies more on content than on vocabulary, phrasing, and structure, and much of the style is acquired by reading existing records.

Health records documentation differs in content, style and structure depending on the situation and the purpose of the note. For instance, daily notes are written by several clinical professionals such as nurses, physiotherapists and physicians, to report on the patient's progression, for internal use by the health care team in the daily care. Other parts of the records, such as radiology reports and discharge notes, are addressed to physicians in other departments of the hospital or to the patients' general practitioner, and are commonly more well-structured and written to summarize impressions, progression or directions/recommendations for further care planning. Linguistic and structural differences in Swedish radiology reports and daily notes have earlier been investigated as a study of genres (Kvist & Velupillai 2014, Smith et al. 2014).

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

Language	Diagnosis 1	Diagnosis 2
Latin	Infarctus myocardii acutus	Encephalitis viralis
English ICD-1	Acute Myocardial infarction	Viral encephalitis
English common	Acute heart attack	Viral brain inflammation
Swedish ICD-10	Akut hjärtinfarkt	Virusencefalit
Swedish common	Akut hjärtattack	Viral hjärninflammation
German ICD-10	Akuter Myokardinfarkt	Virusenzephalitis
French ICD-10	Infarctus aigu du myocarde	Encéphalite virale
Spanish ICD-10	Infarto agudo del miocardio	Encefalitis viral

Table 1. Terms for two diagnoses in Latin and according to ICD-10 in different languages, and corresponding expressions in general English and Swedish.

2.2 Swedification of medical terminology

There are a number of international medical vocabularies also available in Swedish – ICD-10,[1] MeSH,[2] SNOMED CT[3] – developed for standardization purposes in the medical domain, and for maintaining guidelines for terminology usage, including preferred spellings of clinical and medical terms. However, there is little overlap in the actual terminology use in the narrative parts of clinical texts and the terms in these terminologies (Skeppstedt, Kvist & Dalianis 2012). Similar findings have been shown for the text from a medical scientific corpus (Kokkinakis 2011a) and from public health portals (Kokkinakis 2011b).

2.2.1 Latin and Greek in medical terminology

As mentioned above, a considerable part of the medical terminology originates from Latin and Greek (Baney 1948). As with most scientific writing in the 18th century, medical patient records were originally written in Latin (Nilsson 2007), thereby being internationally comprehensible. Different languages have adapted medical terms differently (Van Hoof 1998, Bretschneider et al. 2013). Table 1 shows two examples of the way in which Latin terms have either been adapted or correspond to different terms in English, Swedish, German, French and Spanish. In Swedish, the Latin expressions for diagnoses were used for classification of disorders until 1987. Today, the Swedish medical expressions are used in the ICD-10 terminology, but the Latin expressions are included and kept as a subtitle. Table 2 summarizes Latin and Greek affixes that are common in the medical domain, obtained from Fogelberg & Petersson (2013).

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

Latin prefixes	Latin suffixes	Greek prefixes		Greek suffixes
acu:aku	cida:cid	anthrop:antrop	cam:kam	ectomia:ektomi
circum:cirkum	bilis:bil	arthr:artr	thanat:tanat	genesis:genes
con:kon	formis:form	blephar:blefar	thel:tel	graphia:grafi
saept:sept	ides:is	cardi:kardi	toc:tok	haemia:hemi
prae:pre	formis:form	chol:kol	acust:akust	hexia:hexi
	ilis:il	chole:kole	cephal:cefal	exia:exi
	alis:al	chondr:kondr	copr:kopr	lasis:las
	aris:ar	chrom:krom	dacry:dakry	iatria:iatri
	arius:ari	chromat:kromat	path:pat	ismus:ism
	inus:in	colpo:kolpo	phac:fak	icus:isk
	itis:it	cor:kor	aesthes:estes	lysis:lys
	ivus:iv	core:kore	brachy:braky	mania:mani
	idus:id	dactyl:daktyl	chylo:kylo	odynia:odyni
	osus:os	encephal:encefal	crypt:krypt	oma:om
	lentus:len	galact:galakt	cycl:cykl	osis:os
		gnath:gnat	glyc:glyk	pathia:pati
		gynaec:gynek	macr:makr	algia:algi
		haem:hem	micr:mikr	asthenia:asteni
		haemat:hemat	necr:nekr	atresia:atresi
		lith:lit	pachy:paky	centesis:centes
		morph:morf	phag:fag	desis:des
		myco:myko	phlog:flog	ectasia:ektasi
		nephr:nefr	phono:fono	philia:fili
		onco:onkofal	phos:fos	phobia:fobi
		onych:onyk	photo:foto	plasia:plasi
		ophor:ofor	phys:fys	plegia:plegi
		ophthalm:oftalm	scler:skler	pnoea:pné
		orchi:orki	therm:term	poiesis:poies
		paed:ped	toxic:toxik	ptosis:ptos
		phleb:fleb	troph:trof	ptysis:ptys
		proct:prokt	cac:cak	rhagia:ragi
		psych:psyk	caco:kako	rhexia:rexi
		rhin:rin	cata:kata	rhoea:ré
		sarc:sark	ortho:orto	schisis:skis
		sthen:sten	haemato:hemato	scopia:skopi
				stasis:stas
				stomia:stomi

Table 2. Affix pairs used in this study (original:Swedified), obtained from Nyman (2013b, c, d).

2.2.2 Swedification

The Swedish medical terminology underwent a Swedification of diagnostic expressions in the 1987 update of the Swedish version of the ICD (Smedby 1991).

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

Original → Transliteration	
ae → e	c → k
oe → e	cc → ck
ph → f	ch → k
rh → r	sc → sk
th → t	u → v [after q and ng]

Table 3. Transliteration rules according to the 1987 spelling reform.

The Swedish National Board of Health and Welfare decided to partly change the terms of traditional Latin- and Greek-rooted words. This included a Swedification of Latin and Greek affixes as well as abandoning the original rules for inflections. The purpose of this was to bring the classification language up to date and mirror the contemporary medical language.

The ambition was originally to go even further in the change of expressions and use the translated or genuine older Swedish expressions. However, it was concluded that a more radical change into Swedish terms would not gain acceptance in the medical professional community and the committee for the Swedish ICD classification settled on a degree of Swedification that would be accepted and used.

2.2.3 Spelling reform

The Swedification of diagnostic terms in 1987 was paralleled by a spelling reform in the Swedish ICD classification. However, it took a few years before the Language Committee of the Swedish Medical Association concurred with these recommendations. The spelling reform affected the Swedified versions of medical terminology expressions, while the original Latin expressions, for example involving diagnoses, anatomical structures or microbiological pathogens, kept the classical Latin spelling. The spelling reform aimed for a spelling compatible with the Swedish spelling rules. In this spelling reform, *c* and *ch* pronounced as *k* was changed to *k*, *ph* was changed to *f*, *th* to *t*, and *oe* was changed to *e*, see Table 3. For example, the technical term for *cholecystitis* (inflammation of the gallbladder) is now correctly spelled *kolecystit*, and *oesophagus* is spelled *esofagus*.

According to clinical terminology practice, the author can choose to write a term in either the original multi-word Latin expression, or the Swedified form, and should spell the term accordingly. Thus, the Swedification process does not apply to foreign affixes used in multi-word expressions for anatomical structures

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

(e.g. *musculus tibialis posterior*, *sinus cavernosus*), microbiological pathogens (e.g. *Staphylococcus aureus*) or diagnostic terms (e.g. *amaurosis fugax*, *status epilepticus*).

However, the medical community seems to be a conservative group, and the adherence to the spelling rules in clinical practice has been gradual. Furthermore, because the medical literature is predominantly English nowadays, physicians increasingly get exposed to the English spelling of Latin and Greek words rather than the recommended Swedish one. The English medical language has, like many other languages, kept more of the original Latin spelling in medical expressions than the Swedish has. This has in practice resulted in a multitude of alternate spellings of medical terms in Swedish clinical notes. For example, *tachycardia* (rapid heart) is correctly spelled *takykardi* in Swedish, but is also frequently found as *tachycardi*, *tachykardi*, and *takycardi* (Kvist et al. 2011). The phenomenon of Greek- and Latin-rooted words introducing unusual inflection forms has also been observed in German clinical texts. These words were often used interchangeably with the corresponding German word (Bretschneider et al. 2013).

3. METHODOLOGY FOR DETECTING AFFIX USE IN CLINICAL TEXTS

3.1 Methodological process

For the purpose of providing the statistics of prefix and suffix usage in Swedish clinical texts we use the following processing scheme:

1. Data extraction: token frequency lists
2. Affix string matching
   a. Direct string matching + compound splitting
      i. initial and non-initial prefixes of words
      ii. suffixes as word endings
   b. Pairwise-combinations + compound splitting
      i. initial and non-initial prefixes of words
      ii. suffixes as word endings
3. Expert annotation
4. Result calculation

3.1.1 Data extraction: Token frequency lists

Because of the extremely sensitive nature of the content in the Stockholm EPR corpus, the corpus was tokenized and converted to frequency lists, one for the whole corpus, and one for each subcorpus. Similarly, the comparable corpora were converted to

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

Word sample set	String matching + compound splitting	Pairwise affix matching + compound splitting
colit	**colit**	**colit**
colitis	**colitis**	**colitis**
collit	**collit**	collit
folliculit	**folliculit**	**folliculit**
folliculitis	**folliculitis**	**folliculitis**
myelit	**myelit**	myelit

Table 4. Examples of suffix *-itis* and *-it* detection with two methods: string matching + compound splitting (second column) and pairwise affix matching + compound splitting (third column). Bold font indicates what words were detected by each method.

frequency lists. These lists served as the main sources for the study described in this paper. Section 4.1 below describes the tokenization of the corpora.

3.1.2 Affix string matching

We employed substring matching for finding affixes. Prefix matching has two constraints: initial prefixes that are used at the beginning of words (e.g. ***arthr****osknä*, ***krypt****ococcos*, ***orto****peden*), and non-initial prefixes that are used as succeeding prefixes and/or prefixes in compounds (e.g. *fiber**rhin**oskopi*, *hjärthyper**trof**i*, *elektro**fys**iologisk*). Suffix matching is restricted to the endings of words only (e.g. *polymy**algi***, *acid**osis***, *vir**os***).

Two processing alternatives to affix detection were employed: direct (naïve) string matching + compound splitting and pairwise affix matching + compound splitting. Table 4 illustrates what words containing suffixes *-itis* and *-it* are detected in a word sample set by using these two methods.

Direct string matching results in high recall, i.e. it will guarantee that all affix instances are found, but it may result in a substantial proportion of false positives, i.e. instances erroneously recognized as containing an affix, for instance due to violated morpheme boundaries (e.g. *diureti**kam**edicin*, *överarmsmu**sklern**as*).

In order to reduce the amount of potential false positives, i.e. an attempt to ensure that the identified words contain actual Latin and Greek affixes, we searched for the pairwise combinations in words of original and Swedified affixes. That is, we limit the substring matching to words that occur with both: original and Swedified affixes, e.g. ***haem****atom* and ***hem****atom*. The pairwise-combination matching strategy narrows the observed space of the affix usage by excluding individual words that contain an original or Swedified affix only. In other words, this method means that misspelled variants and/or individual occurrences of one or the other affix type, are excluded in

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

the search space, but it ensures that the detected word pairs contains the exact same word with one Swedified and one original affix.

Additionally, the violation of morpheme boundaries can be improved by compound splitting. For instance, compare two cases: *sensor**i**+**n**euralt* and *fiber+**rin**oskopi*. The compound splitting that we employ in this study is based on using a large general language Swedish dictionary (The NST Dictionary 2007) and a medical domain dictionary, resulting in a precision of 83.5%, and is described in more detail in Grigonytė et al. (2014).

3.1.3 Expert annotation

The final methodological step employed in this study is a manual review of the resulting word pairs containing original and Swedified affixes. In this step, a senior physician manually reviewed the resulting word pairs to identify false positive affix matches such as ***Con**go* (country), ***myc**ket* (Swe: much), *Ka**rin**a* (name), ***kor**tet* (Swe: the card) – which are regular Swedish words and names, not Latin or Greek – and to identify other potential errors, as well as qualitatively categorize and analyse the results. Due to the time costs involved in manual analysis, this step was only employed on the pairwise-combinations (Sections 5.2–5.6), not on the results obtained after employing the direct matching technique (Section 5.1).

This three-step semi-automatic procedure aims at gaining as high quality of the words containing affixes as possible. Alternatively it could be viable to use an entirely automatic procedure by for instance exchanging the manual inspection with an unsupervised morphological segmenter. The state of the art as known from the Morpho Challenge 2010 (Kurimo et al. 2010) has reported the following highest performance for unsupervised segmenters: $F = 64.55\%$ for general English and $F = 47.64\%$ for general German. To our knowledge these segmenters have not been tested on domain data and therefore we can only hypothetically predict that the expected performance for Swedish clinical data would not be better.

3.1.4 Result calculation

Results on affix usage in Swedish clinical texts are calculated by absolute and normalized proportion values. By absolute we mean that statistics is built upon the absolute numbers of occurrences. For the interpretation of these values, especially with the pairwise-combination method, it is necessary to be aware of the effect of infrequent word pairs being overshadowed by one or several very frequent cases. In order to counteract this effect we also use normalized values by type. This way each word pair in the specific affix group is normalized to have an equal proportion of impact. One- and two-sample z-tests of proportions ($p < .0001$, two-tailed) are calculated for statistical significance testing.

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

3.2 Experiments

Nyman (2013b, c, d) lists Latin and Greek affixes that are commonly used in the Swedish medical domain. We select the subset of those affixes for which the Swedification rules apply (Table 2 above) and analyse their usage in Swedish clinical text. We conduct experiments for six distinct affix usage patterns.

Two experiments compare clinical affix usage in the notes of Swedish Electronic Health Records (EHR) with two other medical genres (medical publications and medical online forum articles):

- the proportion of original and Swedified affixes and how it compares to the two other medical genres, and
- the difference, if any, in the use of Latin and Greek affixes in Swedish clinical text compared with the other two genres.

Four experiments are conducted to characterize the usage of affixes in clinical EHR text only: affix usage depending on (3) the position in the word, and (4) the length of the affix. Finally, differences of affix usage between (5) clinical professions and (6) clinical subspecialties are calculated.

4. DATA

4.1 Clinical corpus

The corpus used in this study is the Stockholm Electronic Patient Record (EPR) Corpus with data from the years 2006–2010[4] (Dalianis, Hassel & Velupillai 2009, Dalianis et al. 2012). The corpus contains de-identified patient notes documented in the Electronic Health Records (EHR) system used in Stockholm City Council (TakeCare[5]) with the exception of some categories of records, for example from psychiatry and venereology. In this system, clinical notes are written in semi-structured templates, where each clinical department and profession can define specific templates for their purposes, e.g. a template containing headings such as Past medical history, Current status, Assessment. A template can consist of free-text fields (notes) as well as structured entries such as boxes and dropdown menus with predefined values. The notes are written in Swedish, and by different clinical professionals, e.g. physicians, nurses, dieticians. There is no information about author identity (e.g. names) or other individual distinguishing aspects such as age in the corpus, only information about profession type. The TakeCare EHR system did not supply any support for grammar- or spell-checking during the years 2006–2010.

For this study, only the narrative text was used, leaving out structured parts such as laboratory results and code lists, e.g. diagnosis codes and procedures. All written notes were extracted from the entire document collection[6] and tokenized using an

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

	Corpora of clinical text	Number of types	Number of tokens
	Stockholm EPR Corpus 2006–2010	3,858,107	1,582,329,383
Type	Subcorpora definition		
Profession	physicians	2,806,627	1,015,142,127
	nurses	1,253,662	348,170,835
	assistant nurses	132,296	10,085,551
	physiotherapy practitioners	339,966	49,728,734
	dieticians	157,527	19,049,272
Sub- specialty	Operating specialties	1,088,824	236,144,574
	Oncology	448,843	74,720,211
	Infection	345,079	46,168,146
	Cardiology	374,698	54,057,066
	Neurology	599,146	91,751,642

Table 5. Features of the clinical corpus and its subcorpora (profession 'nurses' includes nurses and midwifes, profession 'assistant nurses' means nurses without academic training, profession 'physiotherapy practitioners' includes physiotherapists, chiropractors, and naprapaths).

adapted version of Stagger (Östling 2013)[7] and word frequency lists were created. Only tokens containing alphabetic characters were used, all converted to lowercase. Although only containing frequency lists from this point, we continue to call this corpus THE STOCKHOLM EPR CORPUS, see Table 5 for details.

4.1.1 Subcorpora

The Stockholm EPR corpus was further divided into two main subcorpora, each with five categories: (i) clinical profession, and (ii) clinical subspecialty (Table 5). Structured data linked to the free text revealed codes for author profession and clinical unit and were used to compile the subcorpora.

The main authors of patient records are physicians and nurses, as can be seen by the corpora sizes in Table 5, but many other clinical professions write progress notes or daily notes of care. Physiotherapists, chiropractors and naprapaths have a common denominator in their focus on anatomical structures and the physiology of the body, and their notes were combined in order to get a sizable corpus. Dieticians have a highly specialized focus of the patients' dietary needs and related pathologies. To study the influence of academic training, corpora were created for notes written both by nurses with an academic education and by assistant nurses (*undersköterskor*) without academic training.

The Stockholm EPR Corpus contains free text written at more than 500 medical units, some of which are within the same hospital department. In order to study differences in language use between clinical subspecialties, several medical units

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

Corpora of medical text	Number of types	Number of tokens
Läkartidningen corpus	442,227	19,588,856
Vårdguiden corpus	46,398	2,810,948

Table 6. Features of medical corpora used for comparison to the clinical corpora.

were combined to form subdomains designed to reflect diverse subspecialties within Karolinska University Hospital, using records for both inpatients and outpatients. A corpus of operating specialties was compiled by pooling the records from several departments of surgery (general surgery as well as plastic, neuro, and thoracic surgery) and orthopaedic surgery. For the other corpora of subspecialties, records were pooled from several wards and outpatient clinics of the respective departments. Most of the authors are physicians and nurses.

4.2 Comparable corpora

For comparison, we also use data from *Läkartidningen* and Vårdguiden. *Läkartidningen* (the journal of the Swedish Medical Association) is a weekly medical scientific and trade-union journal published in Swedish for medical professionals. It contains biomedical scientific publications as well as articles about new medical scientific findings, studies in the pharmaceutical domain, health economic discussions and evaluations, as well as opinion pieces and political discussions. All articles published in *Läkartidningen* are copyrighted, but an openly available corpus containing randomly assembled sentences taken out of context is accessible through Språkbanken[8] (Kokkinakis 2012). The *Läkartidningen* corpus was retrieved in January 2014 from Språkbanken's Korp.[9]

The Vårdguiden corpus contains articles from 1177.se and Vårdguiden.se,[10] which are national Swedish online search engines and medical knowledge repositories dedicated to health related information, services, queries and discussions for the public, provided by all Swedish health care counties and regions. All entries about diseases, facts and recommendations were downloaded from these websites,[11] and tokenized using the adapted version of Stagger. A summary of these corpora are presented in Table 6.

5. RESULTS AND ANALYSIS OF FINDINGS

We present results and analysis from the experiments on the six different affix usage patterns in Swedish clinical text. For each pattern, we summarize the affix matching methodology and the corpora used for the specific pattern analysis.

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

5.1 Pattern 1: Latin and Greek affixes in three medical genres

METHOD: Direct string matching + compound splitting
DATA: The Stockholm EPR Corpus, the *Läkartidningen* corpus, the Vårdguiden corpus

Figures 1 and 2 summarize Latin and Greek suffix and prefix usage in three different corpora. All of the affix pairs occur in the clinical corpus. However some of the affixes do not occur in the comparable corpora, e.g. *circum-* and *cirkum-* or *cac-* and *cak-* in the Vårdguiden corpus. Notably, several (n = 20) affix pairs are not found in the Vårdguiden corpus at all. This latter corpus is written for the general public, not medical professionals with training in medical terminology. Consequently, Swedified forms are more common than original forms, and Latin endings appear only for original Latin multi-word expressions. The proportion of affixes in both original and Swedified forms in term of absolute values is summarized in Figure 1 and Figure 2.

The overall usage of Latin and Greek affixes in original form in all three corpora is low. The majority of affixes are used in Swedified forms. The proportions of original form (Latin and Greek) prefix matches in three corpora are 3.4%, 2.1%, and 1.1%. The online forum genre has the lowest proportion of prefixes in original Latin or Greek form. The proportions of original form (Latin and Greek) suffix matches in three corpora are even lower: 0.8%, 0.5%, and 0.3% respectively.

The observable effect of high proportions of some suffixes and prefixes in original forms in the Vårdguiden and the *Läkartidningen* corpora is due to a very low number of occurrences, e.g. (Greek affixes left, Swedified right):

(1) -poiesis 1 -poies 0
pachy- 3 paky- 0

Another source of complication for interpreting the results of pairwise affixes is that some regular Swedish inflections are similar to foreign suffixes, e.g. the genitive form in multi-word expressions such as *Kaposis sarcom* can be mistaken for *-osis* (in the pair *-osis* and *-os*), as Swedish does not use apostrophes for genitive. Example (2) below includes unwanted pairing shown with the number of occurrences (Greek suffix left, Swedified right). This example illustrates an interesting aspect of the guidelines for the Swedification process – the fact that multi-word Latin expressions should be written in their original form – instances which will not be captured through our chosen methodology.

(2) kaposis 812 kapos 1

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

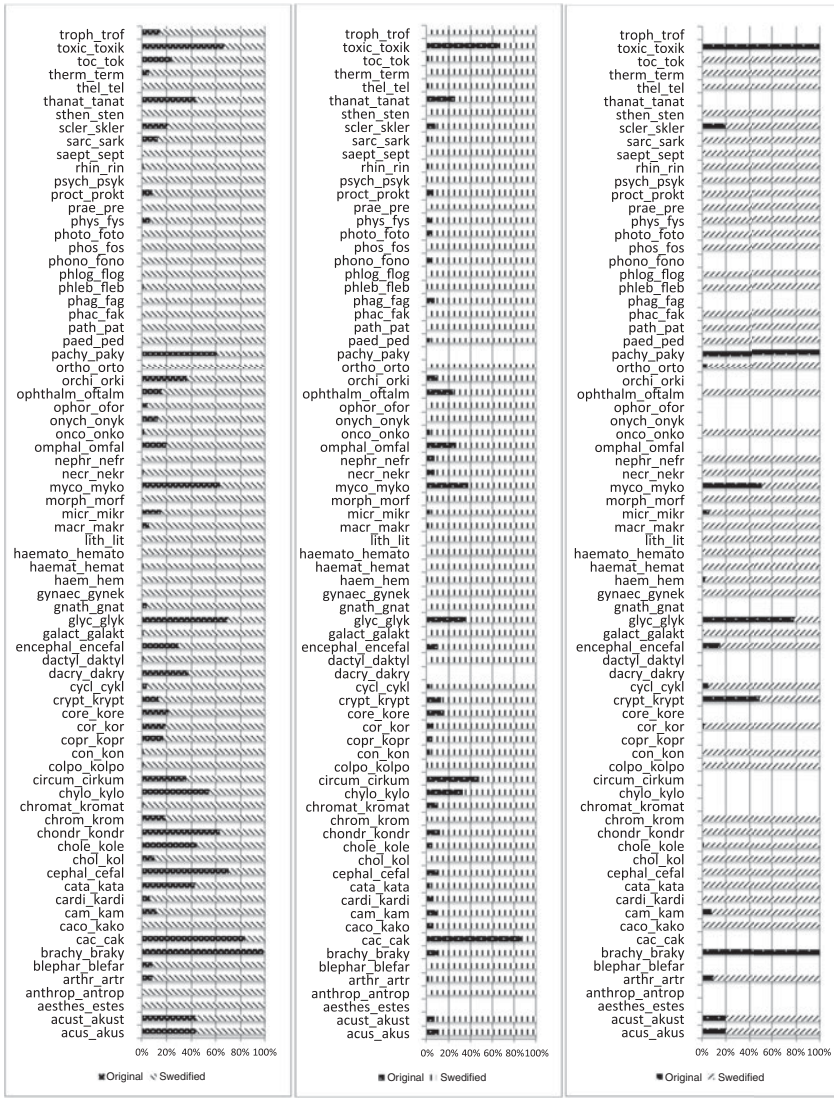

Figure 1. Latin and Greek prefix usage in three genres: clinical text (the Stockholm EPR corpus), scientific articles (the *Läkartidningen* corpus), and medical online information articles (the Vårdguiden corpus).

5.2 Pattern 2: Differences of the usage between Latin and Greek affixes

METHOD: Pairwise-combinations + compound splitting

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

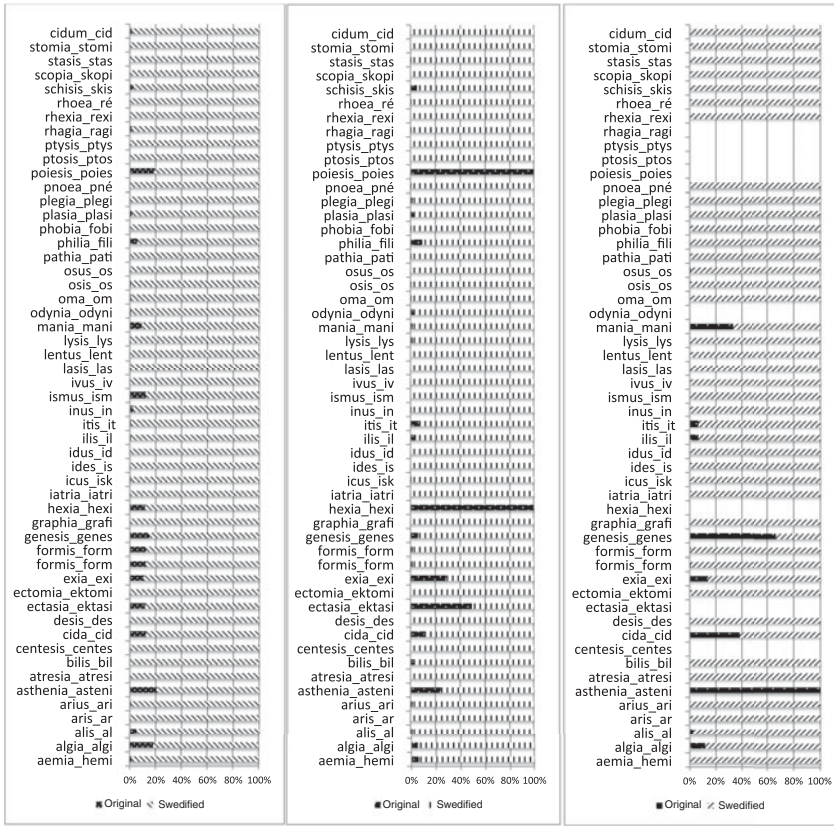

Figure 2. Latin and Greek suffix usage in three genres: clinical text (the Stockholm EPR corpus), scientific articles (the *Läkartidningen* corpus), and medical online information articles (the Vårdguiden corpus).

DATA: The Stockholm EPR Corpus, the *Läkartidningen* corpus, the Vårdguiden corpus

Results for the difference of usage between Latin and Greek affixes in the three corpora are presented in Table 7. First, the Swedified form irrespectively of the type of affix is strongly preferred to the original form in the Swedish medical domain. These differences are statistically significant at the .001 level according to the 1-sample z-tests of proportions ($p < .0001$, two-tailed) across all three corpora. Secondly, prefixes are more Swedified than suffixes. This result holds for both Latin and Greek in both the EPR and *Läkartidningen* corpora; all the differences are statistically significant at the .001 level according to two-sample z-tests of proportions ($p < .0001$, two-tailed). In the smaller Vårdguiden corpus, where many affix pairs are not present,

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

			Stockholm EPR Corpus		*Läkartidningen*		Vårdguiden	
			Part	Occurrences	Part	Occurrences	Part	Occurrences
Latin	Prefix	Found	5/5	—	5/5	—	1/5	—
		Original	0.08	0.38×10^6	0.12	504	0.23	17
		Swedified	0.92	4.33×0^6	0.88	3789	0.77	58
	Suffix	Found	15/15	—	13/15	—	4/15	—
		Original	0.03	$0{,}39 \times 10^6$	0.10	1151	0.2	19
		Swedified	0.97	14.07×10^6	0.90	10249	0.8	74
Greek	Prefix	Found	71/71	—	45/71	—	10/71	—
		Original	0.13	1.15×10^6	0.13	1262	0.15	54
		Swedified	0.87	7.8×10^6	0.87	8613	0.85	299
	Suffix	Found	32/37	—	16/37	—	3/37	—
		Original	0.02	0.09×10^6	0.10	645	0.12	37
		Swedified	0.98	4.79×10^6	0.90	5922	0.88	265

Table 7. The proportions of original and Swedified forms of Latin and Greek affixes in the Stockholm EPR, *Läkartidningen* and Vårdguiden corpora. The statistics are calculated from absolute numbers of occurrences.

there are no significant differences at the .05 level. Thirdly, in the EPR corpus Latin prefixes are more Swedified than Greek prefixes, whereas Greek suffixes are more Swedified than Latin suffixes. The differences are again statistically significant at the .001 level according to two-sample z-tests of proportions ($p < .0001$, two-tailed). On the other hand, there are no significant differences of this kind at the .05 level in the *Läkartidningen* and Vårdguiden corpora. Fourthly, both the prefixes and suffixes of the EPR corpus are more Swedified than in the *Läkartidningen* corpus, which are in turn more Swedified than in the Vårdguiden corpus (note, however, again that many affix pairs are not present in the Vårdguiden corpus). The only exception to this concerns Greek suffixes, where the differences are not significant at the .05 level; the other differences are statistically significant at the .001 level according to two-sample z-tests of proportions ($p < .0001$, two-tailed).

The Swedified version of the adjectival Latin suffixes can be inflected and thus would not be captured by the pairwise-combinations method. For instance, *infraorbitalis* 'infraorbital' (meaning 'located below the eye socket') can in Swedish be inflected as *infraorbital* as well as *infraorbitalt*, depending on the head word (agreement). Two examples of pairwise combination of adjectives are shown with the number of occurrences (Latin suffix left, Swedified right):

(3) infraorbitalis 4414 infraorbital 83
periorbitalis 1 periorbital 1452

In some cases word misspellings can have an influence. Consider two examples of misspelings (*acustisk* and *akusticus*) found in a pairwise combinations originating from the suffix *-icus,* where the correct Swedish spelling would be *akustisk* (Greek left, Swedified right):

(4) acusticus 1080 acustisk 4
akusticus 36 akustisk 1807

Another source of errors comes from abbreviations, for instance *mobil* and *mobilis*, both abbreviations for *mobiliserad,* yielding pairwise combinations (Latin suffix left, Swedified right):

(5) frimobilis 1 frimobil 1
svårmobilis 2 svårmobil 1

The latter type of affix matching error is observed more often with words that appear to contain a Swedified suffix.

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

Original prefix	Number of occurrences	%	Swedified prefix	Number of occurrences	%
cryptokockinfektion	3	0.12	***krypt***okockinfektion	23	0.88
thermotest	69	0.32	***term***otest	151	0.69
mycobacterier	120	0.82	***myk***obacterier	27	0.18
sclerae	2929	0.77	***skler***ae	876	0.23
spond***arth***rit	32	0.01	spond***art***rit	4741	0.99
oeso***phag***it	688	0.73	oeso***fag***it	251	0.27
hypo***therm***ibehandlas	4	0.06	hypo***term***ibehandlas	65	0.94
mikroangio***path***i	59	0.03	mikroangio***pat***i	1648	0.97
meta***plasia***	27	0.01	meta***plasi***	2445	0.99
agora***phobia***	2	0.001	agora***fobi***	1616	0.99
derma***tosis***	60	0.06	derma***tos***	867	0.94
fibr***oma***	152	0.05	fibr***om***	2968	0.95

Table 8. Examples of prefixes and suffixes in different positions of words.

5.3 Pattern 3: Affix usage depending on the position in a word

METHOD: Pairwise-combinations + compound splitting
DATA**:** The Stockholm EPR Corpus

In this section we analyse the impact of the position of an affix in a word. We compare prefixes that occur as the first syllable of a word, prefixes that occur as the second or later syllable of a word and suffixes – the last syllable of a word, see Table 8 for examples.

Figures 3a and 4a illustrates which original and Swedified prefixes and suffixes are found in the clinical corpus. These proportions are based on the normalized values by type. Part (a) of Figure 3 displays the percentage for each suffix found in its original or Swedified form. *Brachy-* for instance is mainly found in its original form, whereas the Swedified *makr-* is preferred to the Greek *macr-*. Part (b) of the figure displays proportions for the same original-Swedified prefix pairs when prefixes are found in the non-initial position of a word. The most prominent changes are observed with for instance *galact–galakt* or *aesthes–estes*. Part (c) of the figure shows the difference of those changes (increase on the positive axis, decrease on the negative) in percentage for each prefix pair.

Figure 4 presents proportions of the found suffixes in normalized values by type and absolute values. The proportion of the Swedified suffix is dominant for most of the suffixes as expressed in absolute values. The suffix graph of the normalized values shows that many not so frequent word types in fact contain original suffixes.

When analysing the pattern of the affix position in a word, we look at initial and non-initial prefixes and suffixes as somewhat 'equal' in an abstract way. By doing

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

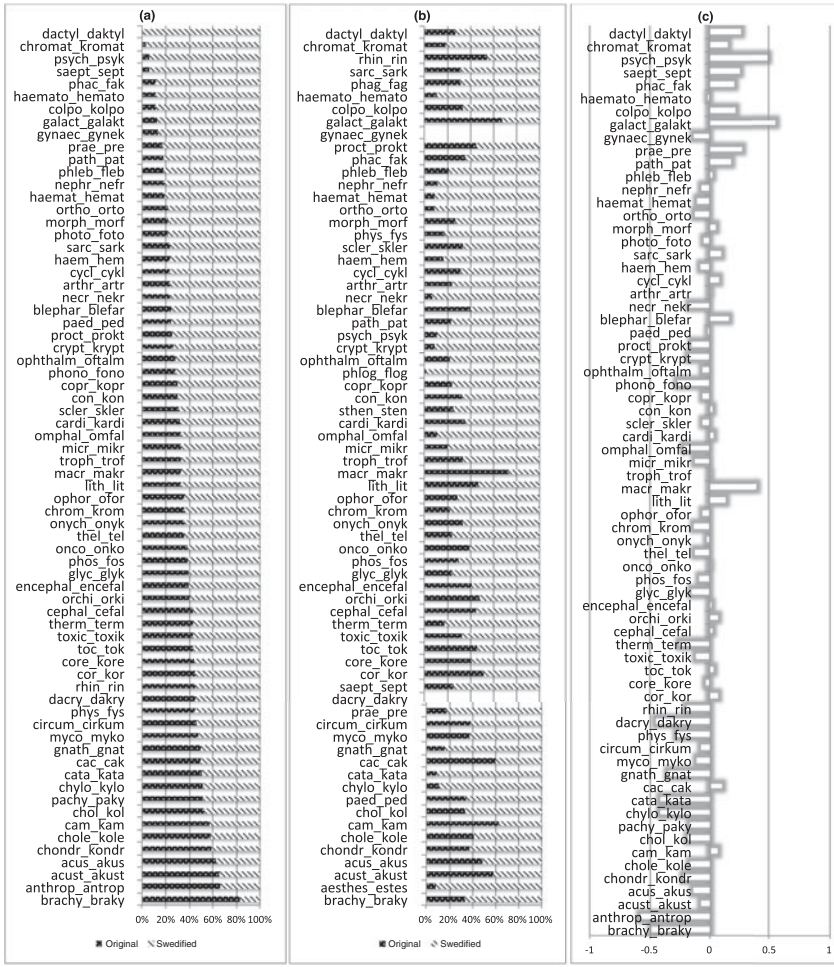

Figure 3. Latin and Greek prefixes found in initial and non-initial positions of words: (a) prefixes found as the initial syllable of a word; (b) prefixes found as the non-initial syllable of a word; (c) the difference between the two. Negative bar means decrease, positive bar – increase.

this we aim at quantitatively identifying whether the position in the word determines how likely the affix is going to be used in its original or Swedified form. Table 9 summarizes our findings in terms of paired words containing original and Swedified affixes.

The findings show that the position in the word does matter for the chance of the affix being used in original or Swedified form. The normalized data reveal that the initial prefixes are found in original form in 34%, non-initial prefixes in 30%, and suffixes in 23% of cases.

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

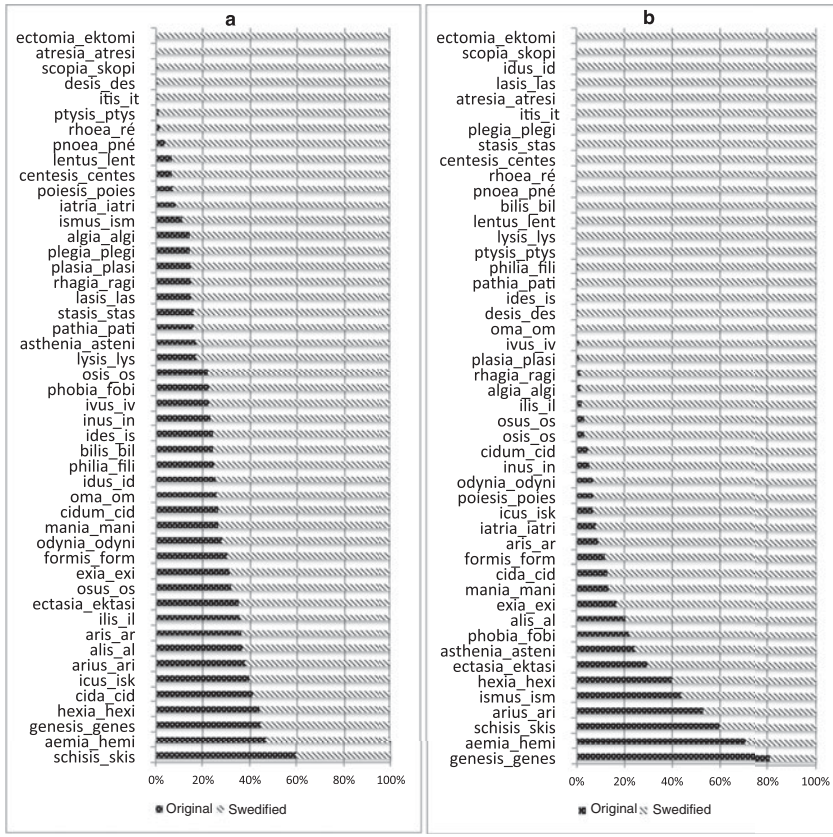

Figure 4. Latin and Greek suffixes found as the proportion of original and Swedified suffixes based on normalized (a) and absolute (b) values.

5.4 Pattern 4: Latin and Greek affix use depending on the length of the affix

METHOD: Pairwise-combinations + compound splitting
DATA: The Stockholm EPR Corpus

This part of the analysis is motivated by the fact that all affixes become either shorter or keep the same number of characters after the Swedification rules apply. Our initial hypothesis is that using the shorter affixes would result in shorter words, which might be important for saving time when clinical notes are composed. Several studies have described a high prevalence of abbreviations in clinical texts, which support the notion that shorter is better in the clinical domain (Xu, Stetson & Friedman 2007, Kvist & Velupillai 2014).

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

	Initial prefix		Non-initial prefix		Suffix	
	Orig	Swe	Orig	Swe	Orig	Swe
Absolute	0.08	0.92	0.20	0.80	0.03	0.97
Normalized	0.34	0.66	0.30	0.70	0.23	0.77

Table 9. Proportions of original (Orig) and Swedified (Swe) affix positions in a word.

	[2–3]		[4]		[5–7]	
Affix length, characters:	Orig	Swe	Orig	Swe	Orig	Swe
Suffix,						
normalized	0.24	0.76	0.25	0.75	0.15	0.85
absolute	0.03	0.97	0.01	0.99	0.01	0.99
Prefix,						
normalized	0.37	0.63	0.30	0.70	0.33	0.67
absolute	0.14	0.86	0.12	0.88	0.20	0.80

Table 10. Proportions of original (Orig) and Swedified (Swe) affixes found depending on the length of the affix.

We have split suffixes and prefixes into three groups according to the length (number of characters) of the Swedified affix: 2–3, 4 and 5–7 characters. Table 10 summarizes the usage patterns depending on the length of the affix.

Our findings did not show any correlation depending on the length of the affix. Both the normalized and the absolute values for suffixes and prefixes show no linear dependence related to the affix length. For suffixes however, we can observe a tendency that when suffixes are very long (5–7) the proportion of them becoming Swedified is larger, suggesting that the shorter ending is preferred.

5.5 Pattern 5: Latin and Greek affix by clinical profession

METHOD: Pairwise-combinations + compound splitting
DATA: The Stockholm EPR Corpus

This section analyses how Latin and Greek affixes are used among five clinical professions: physicians, nurses, assistant nurses, physiotherapy practitioners, and dieticians. Table 11 summarizes the proportions of original and Swedified affixes for the five professions.

The most prominent pattern is that assistant nurses and dieticians proportionally use more original form prefixes than other professions. In terms of normalized values

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

	Physicians		Nurses		Assistant nurses		Physio-therapy specialists		Dieticians	
	Orig	Swe	Orig	Swe	Orig	Swe	Orig	Swe	Orig	Swe
Prefix,										
normalized	0.32	0.68	0.40	0.60	0.40	0.60	0.35	0.65	0.35	0.65
absolute	0.13	0.87	0.11	0.89	0.27	0.73	0.08	0.92	0.28	0.72
Suffix,										
normalized	0.23	0.77	0.34	0.66	0.40	0.60	0.30	0.70	0.25	0.75
absolute	0.04	0.96	0.06	0.94	0.03	0.97	0.11	0.89	0.01	0.99

Table 11. Proportions of original (Orig) and Swedified (Swe) affix found in subcorpora of clinical professions from the Stockholm EPR corpus.

the most conservative groups of professions are nurses and assistant nurses: 40% and 40% of prefixes and 34% and 40% of suffixes are used in the original Latin and Greek form. Especially for suffixes this is a strong contrast to physicians, i.e. 23% of suffixes are used in the original form.

We interpret it as an effect of two factors: the size of the subcorpus and the language differences among the professions. The language of physicians is packed with domain terminology and abbreviations that are ambiguous, for instance '*c*' can mean *cancer*, *cell*, *corpus*, *circa*, and adjective *central*. The absence of pronouns and verbs is yet another very typical feature (Temnikova et al. 2013, Smith et al. 2014). The assistant nurses do not have the same academic training as the physicians, which suggests smaller domain vocabulary and the need to express the same concepts in general and thus more verbose language.

5.6 Pattern 6: Latin and Greek affix by clinical subspecialty

METHOD: Pairwise-combinations + compound splitting
DATA: The Stockholm EPR Corpus

In this section, we present an analysis of how Latin and Greek affixes are used among five clinical subdomains: operating specialty, oncology, infection, cardiology, and neurology. Table 12 summarizes the proportions of original and Swedified affixes for each of the five clinical subspecialties.

We did not find any strong correlation from the statistics presented in Table 13 related to the professions. In terms of absolute affixes found, the most conservative spelling is within the specialties of oncology and infection. In terms of normalized values, prefixes and suffixes are found in rather similar proportions for all specialties.

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

	Surgery		Cardio		Onco		Neuro		Infection	
	Orig	Swe	Orig	Swe	Orig	Swe	Orig	Swe	Orig	Swe
Prefix,										
normalized	0.32	0.68	0.36	0.64	0.36	0.64	0.31	0.69	0.34	0.66
absolute	0.10	0.90	0.10	0.90	0.18	0.82	0.09	0.91	0.16	0.84
Suffix,										
normalized	0.32	0.68	0.33	0.67	0.26	0.74	0.25	0.75	0.33	0.67
absolute	0.06	0.94	0.32	0.68	0.04	0.96	0.09	0.91	0.06	0.94

Table 12. Proportions of original (Orig) and Swedified (Swe) affix found, depending on clinical subspecialty from the Stockholm EPR corpus.

Medical specialty	Prefix	Absolute number of occurrences	Normalized proportion by types
Cardiology	cardi-	9158 (0.21)	0.29
	kardi-	35479 (0.79)	0.71
	cor-	16042 (0.32)	0.57
	kor-	33808 (0.68)	0.43
Surgical specialty	cardi-	3469 (0.23)	0.35
	kardi-	11348 (0.77)	0.65
	cor-	27237 (0.28)	0.51
	kor-	68589 (0.72)	0.49

Table 13. Differences between medical specialties: proportions of original and Swedified prefixes found in the Stockholm EPR corpus.

The detailed manual analysis of the results revealed that the vocabulary is strikingly different between each specialty. For instance, in the cardiology subcorpus, there seems to be more progressive use of *c*/*k* as in *cardi-*/*kardi-* (see Table 13).

We also found a strong lexical preference for some prefixes, like in the case of *cor-*/*kor-* these are very often related with words concerning *coronary* topics (i.e. the vessels in the heart giving angina pectoris or heart attack) but for the surgical specialties it is related to *cortison* treatments and *cortex* or other anatomical structures. In the cardiology subcorpus, we find the Swedified form of the prefix *kor-* as in *koronar-* as the first compound, in contrary to the surgical subcorpus where words related to the *coronary* topic with the original form of the *cor-* suffix over the Swedified *kor-* are preferred. We interpret it as a possible pattern that applies to the specialty specific vocabulary, i.e. terms that are more frequently used within a specialty tend to be more Swedified, whereas the spelling would be more conservative for less frequently used terms in the vocabulary/specialty.

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

6. DISCUSSION

A large proportion of the medical terminology originates from Latin and Greek, in Germanic as well as other languages. In Sweden, since the 1980s, there has been a process of Swedification in the medical domain, which has included a spelling reform and modified affix use. This reform has taken time to have an effect in the medical society.

6.1 Linguistic characterization of Swedish clinical text for knowledge extraction

The present study contributes to the linguistic characterization of Swedish clinical language. Such characterization is essential for constructing automated language analysis tools that can be used for knowledge extraction from clinical text. It has previously been found that many words and expressions in Swedish clinical free text cannot be automatically identified by vocabulary matching to established terminologies (Skeppstedt et al. 2012, Grigonytė et al. 2014). This is in part due to medical jargon and the extensive use of ad hoc abbreviations (Kvist & Velupillai 2014), but also misspellings and foreign words. Also, many words are hybrid words with a spelling being neither Swedish nor Latin or Greek, as a result of the ongoing Swedification and adaptation to new spelling rules. For instance, *bronchit* (contemporary Swedish: *bronkit*) is a common hybrid word found in the clinical corpus, originating from *bronchitis*, losing its suffix but partly keeping an original spelling (*ch* instead of *k*). The findings from this study could be used for development of NLP preprocessing tools that need to be adapted to this domain such as syntactic parsers and part-of-speech taggers (Skeppstedt 2013). For instance, the word pairs of Swedified and original affixes along with the information about proportions resulting from this study can be useful for developing term normalization methods that map term variants to uniform concepts. With sophisticated preprocessing tools, resources such as the Stockholm EPR corpus can be used to build useful applications and systems with the goal to improve health care, such as clinical decision support systems, automatic diagnosis coding (Henriksson, Hassel & Kvist 2011), text simplification for patient empowerment (Grigonytė et al. 2014) and surveillance of adverse events (Tanushi, Kvist & Sparrelid 2014).

6.2 Findings

Both prefixes and suffixes are used in their Swedified form in clinical Swedish text to a very large extent. This pattern remains strong independently of which clinical subcorpus we studied. If contrasted, prefix usage is more conservative than suffix.

As expected, the proportion of Swedified prefixes and suffixes is relatively smaller in clinical texts than in scientific articles and even smaller when compared

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

with medical online information pages. This holds for absolute values and normalized by-type values.

One important factor that would definitely give more insight, but is not covered in this study, is related to misspellings and ad hoc abbreviations, which are abundant in clinical texts, since this type of text is written under time pressure and most often for the purpose of internal healthcare communication. Patient records are seldom corrected after being written. On the other hand, scientific articles and online information pages are reviewed in the process of writing or can even be updated after they have been published, and are written for a broad audience. As an example of a (mis)spelling variation in clinical domain and also demonstrating an obvious need for aggregation of such cases, consider the following pairs with the Greek suffix (left) and the Swedified form (right) (correct spelling in the first pair):

(6)

amaurosis[12] 2958	amauros 811
amorosis 19	amoros 1
amourosis 45	amouros 2
amurosis 44	amuros 1
amaourosis 7	amaouros 1
amarosis 141	amaros 4

To study such examples further with respect to Swedification patterns in clinical, scientific and online health information would require a methodology different than that employed in this study. For instance, terms would need to be normalized and mapped as belonging to the same concept, which would require knowledge about which different variants should be mapped to which concept – within and across corpus types. Moreover, for a deeper study of how these different text types compare in the use of Swedification changes in a larger discourse (that is, not explaining only word pairs), would require also taking context into account.

The difference in the findings for Greek and Latin affixes has shown that Greek prefixes in the original form are more common than Latin in terms of normalized values. The suffix pattern is very similar. It should be noted that the set of Greek affixes used in this study was larger than that of Latin.

The affix analysis depending on the position in a word revealed a positive correlation: initial prefixes are found in larger proportion in their original form if compared with non-initial prefixes and suffixes.

A somewhat surprising finding is that in terms of both the normalized and the absolute values for suffixes and prefixes show no linear dependence related to the affix length, apart from very long suffixes (longer than five characters) for which the proportion of Swedified usage increases.

The analysis of affixes in various sets of subcorpora has shown insignificant differences in affixes found in the different subdomains of clinical text on the basis

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

of surface parameters. After a closer examination, we conclude that the vocabulary in the different subcorpora clearly reflects the divergent working tasks of different professionals and different subspecialties. Lexical features of a subdomain language can be used for unsupervised clustering of text (Patterson & Hurdle 2011, Zeng et al. 2011), but these studies do not specifically focus on the usage of terminology with foreign origin. Patterson & Hurdle (2011) suggest that differences in language use between professionals, which create disjoint sublanguages, influence the creation of NLP tools for clinical text. A tool which relies on term statistics or semantics and is trained on one clinical note type may not work as well on another.

The analysis of the affix use by different healthcare professionals was limited by the methodology of only extracting word pairs. Thus, the higher frequency of original affixes for the assistant nurses (without academic training) may not necessarily reflect a trend of using original Latin/Greek affixes, as we found very few affix pairs for this group of professionals. A possible explanation can be that assistant nurses are unused to write these words, and therefore are unsure of the spelling.

When analysing subcorpora for different medical subspecialties, there are apparent differences in the use of specific expressions within an affix group, as was shown for *cardi-/kardi-* in Section 5.6. There are, on this level, striking differences both in the number of instances found for different expressions and for affixes found for certain expressions. The findings that the cardiology subspecialty uses Swedified prefixes for expressions specific to their line of profession is in contrast to the findings for the surgical specialty, where they are more likely to use the original affix *chol-* for a large number of expressions for gastrointestinal terms, e.g. *cholecystitis* (gallbladder inflammation) instead of the Swedified *kolecystit*.

6.3 Limitations

Although this study is based on the largest existing data set of Swedish clinical text available for research, there are some limitations. The pairwise-combination matching strategy narrows the observed space of the affix usage by excluding individual words containing only an original or Swedified affix without a matching word with the other affix. However, with this strategy, we are able to precisely study their usage IN COMBINATION, and given the very large size of the Stockholm EPR corpus (1.6 billion tokens), we believe that these found combinations reveal a sufficient approximation of Swedification patterns. We intend to further study the number of word types that were missed because of this strategy – word types with either exclusively Latin or Greek spelling, or exclusively Swedified spelling – and analyse whether or not we find additional patterns through this.

In the frame of this study it was not possible to perform a manual review of the results from direct matching (pattern 1 described in Section 5.1 above). It also has to be noted that the state-of-the-art processing tools (like morphological segmenters and

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

part-of-speech taggers) were not applicable in this study because their performance is currently not meeting the required level for this domain. That partially stems from the low lexical coverage as there are no dictionaries that could deal with at least 50% of the vocabulary used in the Stockholm EPR corpus (i.e. almost four million types, whereas the largest Swedish dictionary resource contains 900,000 types, and the largest Swedish medical domain dictionary contains 500,000 terms). In future studies, we intend to extend the manual review analysis to a larger set in order to be able to quantify how well the employed string matching techniques work for this task.

Furthermore, it was not possible to add a time axis as an additional variable in this study, nor information about author age or other characteristics that would have been informative for understanding changes over time or whether or not there are differences in word usage depending on author age. Since the corpus only covers the years 2006–2010, we suspect that changes over time would not be evident for such a relatively short time period, but will investigate if this information could be extracted and further studied in our future work.

7. CONCLUSIONS

This case study has explored the use of Latin and Greek affixes in medical texts of three types; patient records, scientific medical text and online medical information for laymen. Special attention has been given to different domain languages/subdomains of patient records according to profession and medical specialty. The research has been performed on a very large corpus of Swedish clinical text, the Stockholm EPR Corpus, and compared with medical language from *Läkartidningen* and Vårdguiden. By studying pair frequencies of Latin or Greek affixes in original and Swedified form in these corpora, we have been able to obtain precise measures of the usage of these affixes in the Swedish medical domain, and characterize this in more detail. We have conducted experiments using several distinct patterns with the aim of explaining the numerous variations of the usage of Latin and Greek affix that are manifested in Swedish medical text.

The results of this study show that to a large extent affixes in clinical text are Swedified. The Swedification of clinical text is, however, less common when compared with other medical domain genres, such as scientific publications and online medical texts for laymen.

We have observed that prefixes are more likely to be preserved than suffixes. This is also correlated with the quantitative study of the affixes related to the position of the word. This general pattern seems to be consistent with the Swedish word formation practice, where the productivity of suffixes is greater than prefixes in the sense that suffixes are more common in absolute terms than prefixes; perhaps this is

https://doi.org/10.1017/S0332586515000293 Published online by Cambridge University Press

an indication that suffixes are more likely to be Swedified on the grounds that they are more common.

To our knowledge, this is the first study on a systematic characterization and analysis of the behaviour of Latin and Greek affixes in Swedish medical text.

ACKNOWLEDGEMENTS

The authors wish to express their gratitude to Björn Smedby for kindly reviewing and confirming the history of the Swedification process, to Martin Duneld for excellent technical assistance generating Vårdguiden data, and to anonymous reviewers for their comments. We are grateful to Hercules Dalianis for the initiative of Stockholm EPR Corpus. This work was partially funded by the Vårdal Foundation and Swedish Research Council (350-2012-6658), and supported by Swedish Fulbright Commission and the Swedish Foundation for Strategic Research through the project High-Performance Data Mining for Drug Effect Detection (ref. no. IIS11–0053).

NOTES

1. International Classification of Diseases.
2. Medical Subject Headings.
3. Systematized Nomenclature of Medicine – Clinical Terms.
4. This research was approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2012/2028–31/5.
5. http://www.cgm.com/se/index.se.jsp.
6. Timestamp information was not included in this data extract.
7. Adaptations were made in order to correctly tokenize domain specific abbreviations such as ‘pt’ for ‘patient’.
8. Språkbanken, http://spraakbanken.gu.se/korp.
9. http://spraakbanken.gu.se/swe/resurs/.
10. Vårdguiden.se and 1177.se are now joined into the same website: http://www.1177.se/.
11. http://www.vardguiden.se/Sjukdomar-och-rad/Omraden/, collected on 13 September 2013 and http://www.1177.se/Stockholm/Fakta-och-rad/, collected in September 2013.
12. *Amaurosis fugax* (a form of temporary blindness) is only used in original form.

REFERENCES

Allvin, Helen. 2010. Patientjournalen som genre. En text- och genreanalys om patientjournalers relation till patientdatalagen [The patient record as genre: A text and genre analysis of the relationship of patient records and the patient data act]. Bachelor thesis, Department of Nordic Languages, Stockholm University.

Baayen, R. Harald. 2010. The directed compound graph of English: An exploration of lexical connectivity and its processing consequences. In Susan Olson (ed.), *New Impulses in Word-formation* (Linguistische Berichte Sonderheft 17), 383–402. Hamburg: Buske.

Banay, George L. 1948. An introduction to medical terminology I: Greek and Latin derivations. *Bulletin of the Medical Library Association* 36(1), 1–27.

Bretschneider, Claudia, Sonja Zillner & Matthias Hammon. 2013. Identifying pathological findings in German radiology reports using a syntacto-semantic parsing approach. In Cohen et al. (eds.), 27–35.

Cohen, K. Bretonnel, Dina Demner-Fushman, Sophia Ananiadou, John Pestian & Junichi Tsujii (eds.). 2013. *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP 2013)*, 27–35. Sofia, Bulgaria: Association for Computational Linguistics.

Dalianis, Hercules, Martin Hassel, Aron Henriksson & Maria Skeppstedt. 2012. Stockholm EPR Corpus: A clinical database used to improve health care. In Pierre Nugues (ed.), *Proceedings of the 4th Swedish Language Technology Conference (SLTC)*, Lund, 17–18 October, 25–26.

Dalianis, Hercules, Martin Hassel & Sumithra Velupillai. 2009. The Stockholm EPR Corpus – characteristics and some initial findings. In Peter A. Bath, Göran Petterson & Thomas Steinschaden (eds.), *Proceedings of ISHIMR 2009. Evaluation and Implementation of E-health and Health Information Initiatives: International Perspectives*. 14th International Symposium for Health Information Management Research, Kalmar, Sweden, 243–249.

Fogelberg, Magnus & Göran Petersson (eds.). 2013. *Medicinens språk* [The language of medicine]. Stockholm: Liber.

Friedman, Carol, Pauline Kra & Andrey Rzhetsky. 2002. Two biomedical sublanguages: A description based on the theories of Zellig Harris. *Journal of Biomedical Informatics* 35(4), 222–235.

Grigonytė, Gintarė, Maria Kvist, Sumithra Velupillai & Mats Wirén. 2014. Improving readability of Swedish Electronic Health Records through lexical simplification: First results. In Sandra Williams, Advaith Siddharthan & Ani Nenkova (eds.), *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, Association for Computational Linguistics, 74–83.

Hagège, Caroline, Pierre Marchal, Quentin Gicquel, Stefan Darmoni, Suzanne Pereira & Marie-Hélène Metzger. 2011. Linguistic and temporal processing for discovering hospital acquired infection from patient records. In David Riano, Annete ten Teije, Silvia Miksch & Mor Peleg (eds.), *Proceedings of the ECAI 2010 Conference on Knowledge Representation for Healthcare, KR4HC'10*, 70–84. Berlin & Heidelberg: Springer.

Hay, Jennifer B. & Ingo Plag. 2004. What constrains possible suffix combinations? On the interaction of grammatical and processing restrictions in derivational morphology. *Natural Language & Linguistic Theory* 22, 565–596.

Henriksson, Aron, Martin Hassel & Maria Kvist. 2011. Diagnosis code assignment support using random indexing of patient records – a qualitative feasibility study. In Mor Peleg, Nada Lavrac & Carlo Combi (eds.), *Proceedings of the 13th Conference on Artificial Intelligence in Medicine (AIME)*, Bled, Slovenia, 348–352.

Kokkinakis, Dimitrios. 2011a. What is the coverage of SNOMED CT on scientific medical corpora? In Anne Moen, Stig Kjær Andersen, Jos Aarts & Petter Hurlen (eds.), *Procedings of XXIII International Conference of the European Federation for Medical Informatics*, 814–818. Amsterdam: IOS Press.

Kokkinakis, Dimitrios. 2011b. Evaluating the coverage of three controlled health vocabularies with focus on findings, signs and symptoms. In Bolette Sandford Pedersen,

Gunta Nešpore & Inguna Skadiņa (eds.), *The 18th Nordic Conference of Computational Linguistics (NODALIDA)* (NEALT Proceedings Series12), 27–31.

Kokkinakis, Dimitrios. 2012. The journal of the Swedish Medical Association – a corpus resource for biomedical text mining in Swedish. In Sophia Ananiadou, K. Bretonnel Cohen, Dina Demner-Fushman & Paul Thompson (eds.), *The Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM), an LREC Workshop*, Turkey, 40–44.

Kurimo, Mikko, Sami Virpioja, Ville Turunen & Krista Lagus. 2010. Morpho Challenge Competition 2005–2010: Evaluations and results. *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, Association for Computational Linguistics, 87–95.

Kvist, Maria, Maria Skeppstedt, Sumithra Velupillai & Hercules Dalianis. 2011. Modeling human comprehension of Swedish medical records for intelligent access and summarization systems – future vision, a physician's perspective. In Rune Fensli & Jan Gunnar Dale (eds.), *Proceedings of Scandinavian Health Informatics Meeting*, 31–35.

Kvist, Maria & Sumithra Velupillai. 2014. SCAN: A Swedish Clinical Abbreviation Normalizer. Further development and adaptation to radiology. In Evangelos Kanoulas, Mihai Lupu, Paul Clough, Mark Sanderson, Mark Hall, Allan Hanbury & Elaine Toms (eds.), *Proceedings from Conference and Labs of the Evaluation Forum (CLEF 2014)* (Lecture Notes in Computer Science 8685), Sheffield, UK, September 2014, 62–73.

Laippala, Veronika, Filip Ginter, Sampo Pyysalo & Tapio Salakoski. 2009. Towards automated processing of clinical Finnish: Sublanguage analysis and a rule-based parser. *International Journal of Medical Informatics* 78:e7–e12.

Nilsson, Inga. 2007. *Medicinsk dokumentation genom tiderna: En studie av den svenska patientjournalens utveckling under 1700-talet, 1800-talet och 1900-talet* [Medical documentation though time: A study of the Swedish patient record development during the 18th, 19th and 20th century] (Enheten för medicinens historia). Lund: Medical Faculty, Lund University.

NST Dictionary. 2007. Nasjonalbiblioteket. www.nb.no/Tilbud/Forske/Spraakbanken/Tilgjengelege-ressursar/Leksikalske-ressursar (accessed 17 October 2014).

Nyman, Hans. 2013a. Latinet och svenskan [Latin and Swedish]. In Fogelberg & Petersson (eds.), 42–47.

Nyman, Hans. 2013b. Övriga latinska morfem och vanliga fraser [Additional Latin morphemes and common phrases]. In Fogelberg & Petersson (eds.), 100–106.

Nyman, Hans. 2013c. Grekiskan [Greek]. In Fogelberg & Petersson (eds.), 107–120.

Nyman, Hans. 2013d. Prefix [Prefixes]. In Fogelberg & Petersson (eds.), 121–128.

Östling, Robert. 2013. Stagger: An open-source part of speech tagger for Swedish. *Northern European Journal of Language Technology* 3, 1–18. [Linköping: Linköping University Electronic Press]

Patterson, Olga & John F. Hurdle. 2011. Document clustering of clinical narratives: A systematic study of clinical sublanguages. *American Medical Informatics Association Annual Proceedings 2011,* 1099–1107.

Skeppstedt, Maria. 2013. Adapting a parser to clinical text by simple pre-processing rules. In Cohen et al. (eds.), 98–101.

Skeppstedt, Maria, Maria Kvist & Hercules Dalianis. 2012. Rule-based entity recognition and coverage of SNOMED CT in Swedish clinical text. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard,

Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC2012*, 1250–1257.

Smedby, Björn. 1991. Medicinens Språk: språket i sjukdomsklassifikationen – mer konsekvent försvenskning eftersträvas [Language of medicine: The language of diagnose classification. More consistent Swedification sought]. *Läkartidningen* 88(16), 1519–1520.

Smedby, Björn. 2013. Klassifikationer och kodverk [Classifications and coding]. In Fogelberg & Petersson (eds.), 180–189.

Smith, Kelly, Beata Megyesi, Sumithra Velupillai & Maria Kvist. 2014. Professional language in Swedish clinical text: Linguistic characterization and comparative studies. *Nordic Journal of Linguistics* 37(2), 297–323.

Socialdepartementet. 2008. *Patientdatalagen* [Patient data act]. Svensk författningssamling 2008:355, with changes 2014:829.

Surján, György & Gergely Héja. 2003. About the language of Hungarian discharge reports. *Studies in Health Technology and Informatics* 95, 869–873.

Tanushi, Hideyuki, Maria Kvist & Elda Sparrelid. 2014. Detection of healthcare-associated urinary tract infection in Swedish Electronic Health Records. In Manuel Grana, Carlos Toro, Robert J. Howlett & Lakhmi C. Jain (eds.), *Studies in Health Technology and Informatics* 207, 330–339.

Temnikova, Irina P., Ivelina Nikolova, William A. Baumgartner Jr., Galia Angelova & K. Bretonnel Cohen. 2013. Closure properties of Bulgarian clinical text. In Galia Angelova, Kalina Bontcheva & Ruslan Mitkov (eds.), *Recent Advances in Natural Language Processing*, 667–675. Amsterdam: John Benjamins.

Van Hoof, Henri. 1998. The language of medicine: A comparative ministudy of English and French. In Henry Fischbach (ed.), *Translation and medicine*, 49–65. Amsterdam: John Benjamins.

Xu, Hua, Peter D. Stetson & Carol Friedman. 2007. A study of abbreviations in clinical notes. In Jonathan M. Teich, George Hripcsak & Jaap Suermondt (eds.), *American Medical Informatics Association Annual Proceedings 2007*, 821–825.

Zeng, Qing T., Doug Redd, Guy Divita, Samah Jarad, Cynthia Brandt & Jonathan R. Nebeker. 2011. Characterizing clinical text and sublanguage: A case study of the VA clinical notes. *Journal of Health & Medical Informatics* 2011:S3.